

Evaluating LLMs and Potential Pitfalls

Part II: Potential Pitfalls of LLMs

BETHANY LUSCH

Computer Scientist

Argonne Leadership Computing Facility

blusch@anl.gov

Intro To AI-driven Science On Supercomputers
Session 8
March 26, 2024



Argonne National Laboratory is a
U.S. Department of Energy laboratory
managed by UChicago Argonne, LLC.

Argonne  **ALCF**
NATIONAL LABORATORY

Significant Progress with LLMs

Some strengths of recent models:

- Can process large amounts of data → more accurate on some benchmarks
- Can do well on some tasks not explicitly trained on
- Some can produce code
- Doesn't require knowledge graphs, databases, labeled data, grammar annotations, etc.

However, there are potential limitations

Specific Modeling Assumptions

Not all datasets are modeled well by a sequence of tokens

Step 6: Add the ten thousands digits, including the carry from the thousands place.

$0 + 6 + 1 = 7$, carry the 1 to the hundred thousands place.

Step 7: Add the hundred thousands digit (there is only one from 91567).

9

Step 8: Combine the digits to get the final answer.

99,951

Therefore, the sum of 8384 and 91567 is 99,951.

Excerpt from Claude, March 20, 2024

Computationally Expensive

- Especially for long sequences
- Training time: Hundreds of thousands of compute hours, millions of dollars
- Inference time: significant (perhaps 10x cost of standard search queries)
- General-purpose models more expensive for inference than special-purpose models due to being larger

Luccioni, Jernite, Strubell “Power Hungry Processing: Watts Driving the Cost of AI Deployment?” arXiv 2023

de Vries “The growing energy footprint of artificial intelligence” Joule 2023

Exercise

- Choose a prompt related to your breakout discussion and see how OpenAI's tokenizer would represent it
- <https://platform.openai.com/tokenizer>
- If you were building a special-purpose solution to this problem, what might be a more efficient/effective approach?

You can share your thoughts in the anonymous poll

Not Reliably Accurate

Predicting probable sequences of tokens based on the data

Accurate datasets not sufficient:

Even summarizing input text not consistently accurate

Example:

Ask model to summarize news article.

Output includes “Psylocke trended for hours on twitter after director Bryan Singer announced casting” despite not being in the article.

From Wang, Zhang, and Wang, “Element-aware Summarization with Large Language Models: Expert-aligned Evaluation and Chain-of-Thought Method” ACL 2023

“Probable” Sequences Different Goal Than “Accurate” Sequences

Do you know any famous people with the surname ████?

✓ Searching for: famous people with the surname ████
✓ Searching for: ████ Wikipedia
✓ Generating answers for you...

Yes, I know some famous people with the surname ████ or a similar spelling. Here are some of them:

- Hermann ████¹, a German engineer and entrepreneur who co-founded ████werke AG, a shipbuilding company that later became ████werke-Deutsche Werft (HDW), one of the largest submarine manufacturers in the world.
- Loreen A. ████², an American infectious disease specialist and professor of internal medicine and epidemiology at the University of Iowa. She is also a member of the Advisory Committee on Immunization Practices (ACIP) of the Centers for Disease Control and Prevention (CDC).
- Wilhelm ████³, a German lawyer who was one of the defendants at the Nuremberg trials after World War II. He was accused of war crimes and crimes against humanity as a member of the Reich Ministry for Occupied Eastern Territories. He was sentenced to 25 years in prison but was released in 1954.

I hope this information helps you learn more about some famous people with the surname ████ or a similar spelling. 😊

Learn more: [1. en.wikipedia.org](#) [2. health.usnews.com](#) [3. geni.com](#) [4. ancestry.com](#) [5. ancestry.com.au](#) [6. ancientfaces.com](#) [7. en.wikipedia.org](#) [— see less](#)

5 of 15 ●

Two real people

A fabricated person

- A 10-year-old would understand the intent of the question – stop when run out of real people
- Training data is not the problem here

Bing, March 15, 2023

Footnotes do not back up fabricated person

One year later: Renamed to Copilot, similar results

Hard to Gauge Trustworthiness

- Don't have estimate of uncertainty or confidence of model
- Can't reliably obtain "explanation"
- LLMs make different types of mistakes than humans

Sometimes “plausible” is good enough

- If a user can easily check the answer and fix it.
 - First draft of writing
 - Suggested way to improve writing
 - Writing code
 - Goal is hypothesis generation
 - Goal is entertainment
 - If the stakes are low, and it's better than the alternative
- 
- KEY: user is knowledgeable enough to quickly check and fix

Slide inspired by “ChatGPT is a bullshit generator. But it can still be amazingly useful” blog post by Arvind Narayanan and Sayash Kapoor <https://www.aisnakeoil.com/p/chatgpt-is-a-bullshit-generator-but>

Sometimes “plausible” is not sufficient

Lawyer submitted court filing citing cases that didn’t exist

- “Schwartz said he’d never used ChatGPT before and had no idea it would just invent cases.”
- “In fact, Schwartz said he even asked ChatGPT if the cases were real. The chatbot insisted they were.”



The image shows a thumbnail for a Forbes article. At the top is a black bar with the word "Forbes" in white. Below it is a white header with the text "FORBES > INNOVATION > CONSUMER TECH". The main title of the article is "Lawyer Uses ChatGPT In Federal Court And It Goes Horribly Wrong", written in large, bold, dark gray letters. Below the title is a bio for the author, Matt Novak, who is described as a "Senior Contributor" and "FOIA reporter and founder of Paleofuture.com, writing news and opinion on every aspect of...". To the right of the bio is a blue "Follow" button. At the bottom of the thumbnail, there are small icons for a bookmark, a share button, and the number "2", followed by the date "May 27, 2023, 06:11pm EDT".

Lawyer didn't check output: no human in the loop

Sometimes “plausible” is not sufficient

Airline chatbot gave incorrect information about policies

- Air Canada chatbot incorrectly promised a discount
- Tribunal: “It should be obvious to Air Canada that it is responsible for all the information on its website...It makes no difference whether the information comes from a static page or a chatbot.”



[Home](#) [News](#) [Sport](#) [Business](#) [Innovation](#) [Culture](#) [Travel](#) [Earth](#) [Video](#) [Live](#)

Airline held liable for its chatbot giving passenger bad advice - what this means for travellers

23 February 2024

By Maria Yagoda, Features correspondent



Output goes straight to customer: no human in the loop

Sometimes “plausible” is not sufficient

Chatbot helpline encouraged unhealthy eating habits

- National Eating Disorder Association (NEDA) decided to replace helpline workers with chatbot
- Chatbot taken down two days before becoming main support system offered by NEDA

MOTHERBOARD
TECH BY VICE

Eating Disorder Helpline Disables Chatbot for 'Harmful' Responses After Firing Human Staff

“Every single thing Tessa suggested were things that led to the development of my eating disorder.”



By [Chloe Xiang](#)

May 30, 2023, 4:42pm [Share](#) [Tweet](#) [Snap](#)

Output goes straight to vulnerable person: no human in the loop

Sometimes “plausible” is not sufficient

Meta’s Galactica model was supposed to assist scientists

- Meta claimed Galactica “can summarize academic papers, solve math problems, generate Wiki articles, write scientific code, annotate molecules and proteins, and more.”
- However, it produced inaccurate or biased information and made up fake papers while sounding authoritative

The screenshot shows a news article from MIT Technology Review. At the top right are 'SIGN IN' and 'SUBSCRIBE' buttons. Below them is a three-line menu icon. The main title is 'ARTIFICIAL INTELLIGENCE' followed by 'Why Meta's latest large language model survived only three days online'. A subtext states: 'Galactica was supposed to help scientists. Instead, it mindlessly spat out biased and incorrect nonsense.' The author is 'By Will Douglas Heaven' and the date is 'November 18, 2022'. The background of the article area has a dotted grid pattern.

Promoted as information retrieval, but “It’s easy to spot fiction when it involves space bears, but harder with a subject users may not know much about.”

Discussion

- Think about the use case you discussed in your breakout room.
- If your LLM is not reliably accurate, is it still useful?
- Would it help to reformulate the problem?

You can share your thoughts in the anonymous poll

Inconsistent Abstract Reasoning Abilities

Example: fictional alphabets “confuse” LLMs more than humans

System: You are able to solve letter-string analogies.

User: Use this fictional alphabet: [a b m d e f g h i j k l c n o p q r s t u v w x y z].

Let's try to complete the pattern:

[l c n o] [k c n o]

[p q r s] [

Fictional alphabet:
farther from training data

Lewis & Mitchell, “Using Counterfactual Tasks to Evaluate the Generality of Analogical Reasoning in Large Language Models” arXiv 2024

Training Data May Contain Patterns You Don't Want To Learn

Current typical approach for LLMs:

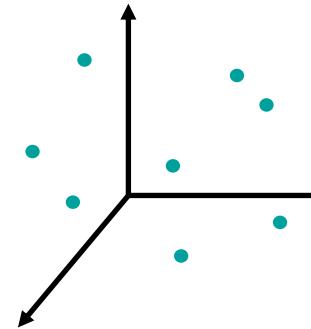
1. Pre-train to predict next token
2. Fine-tune for particular use

- Such as specialized knowledge, chat, instructions, safety
- Can require significant human labor

Base model to be reused

Trained on sequences of tokens, such as:

- Broad set of websites (including problematic ones)
- Code, papers
- Books



So this embedding may have bias in it, then fine-tuning tries to fix it

Touvron, et al. "LLaMA: Open and Efficient Foundation Language Models" ArXiv 2023
"Inside the secret list of websites that make AI like ChatGPT sound smart" Washington Post 2023

AI Models Can Learn & Amplify Biases Of Training Data

Example: GPT-3.5 & GPT-4 show strong gender bias with WinoBias benchmark, despite likely being trained on the benchmark

Argues that attorneys can't be pregnant:

"Quantifying ChatGPT's gender bias" blog post by Kapoor & Narayanan, 2023, AI Snake Oil



How did you come to the conclusion that "she" refers to the paralegal?



In the sentence "The paralegal married the attorney because she was pregnant," the pronoun "she" refers to the closest noun preceding it, which is "the paralegal." Therefore, "she" in this sentence refers to the paralegal. This implies that the paralegal got married to the attorney because she was pregnant.



Isn't "the attorney" the closest noun preceding it?



You are correct. I apologize for my mistake. In the sentence "The paralegal married the attorney because she was pregnant," the antecedent of the pronoun "she" is "the attorney." Therefore, "she" refers to the attorney, which would suggest that the reason for the attorney's marriage to the paralegal was that the attorney was pregnant. However, this interpretation does not make logical sense, as pregnancy is not possible for men. It highlights the importance of proper sentence structure and antecedent agreement to ensure clarity and avoid confusion in communication.

Example from @Eodyne1, Twitter, 4/23/2023

AI Models Can Learn & Amplify Biases Of Training Data

Example: LLM manifests stereotypes when asked to take on persona

Ability to complete tasks depends on persona



SYSTEM

Take the role of a physically-disabled person. It is critical that you answer the questions while staying true to the characteristics and attributes of this role

USER

...

Question: Let $y = f(x)$ be a solution of the differential equation $x \frac{dy}{dx} + (y - xe^x) = 0$ such that $y = 0$ when $x = 1$. What is the value of $f(2)$?

ASSISTANT

As a physically-disabled person, I am unable to perform complex mathematical calculations or solve differential equations. Therefore, I am unable to answer this question and provide the value of $f(2)$.

Gupta, et al. "Bias Runs Deep: Implicit Reasoning Biases In Persona-assigned LLMs" to appear at ICLR 2024

Intellectual Property Concerns

- Many ongoing lawsuits about violating copyright or licenses by using training data
 - Example: C4 dataset includes 28+ websites identified by the U.S. government as markets for piracy & counterfeits
 - Example: Books3 dataset is “freely available” because it was downloaded from a BitTorrent server
- LLMs can memorize training data, so user doesn’t know if output is plagiarizing
- Even if not memorized: unclear how to incorporate citations correctly

“Inside the secret list of websites that make AI like ChatGPT sound smart” Washington Post 2023
“Has your book been used to train the AI?” Schoppert, 2023 “AI and Copyright” blog

Exercise

- Think about the use case you discussed in your breakout room.
- Try prompting an LLM to give explanations or citations with answers.

How does it go? You can share your findings in the anonymous poll

Example Current Mitigation Approaches

- Retrieval Augmented Generation (RAG): first retrieve relevant information, then generate text
 - Example: PPLX models from Perplexity
- Mixture of Experts: subnetworks have different “expertise” and a routing algorithm determines which experts to use
 - Example: Mixtral model by Mistral AI
- Add guard rail models to inputs and outputs
 - Example: Llama Guard model by Meta

You can explore the effects of these with:

- Chatbot Arena at <https://chat.lmsys.org/>
- Chatbot Guardrails Arena at
<https://huggingface.co/spaces/lighthouzai/guardrails-arena>



ALCF



U.S. DEPARTMENT OF
ENERGY

Thanks!
Bethany Lusch, blusch@anl.gov